VU Research Portal

The Riemannian interpretation of Gauss-Newton and Scoring, with application to system identification

Peeters, R.L.M.; Hanzon, B.

1992

Faculteit der Economische Wetenschappen en Econometrie

Serie Research Memoranda

The Riemannian Interpretation of Gauss-Newton and Scoring, with Application to System Identification

Ralf L.M. Peeters
Bernard Hanzon

Research Memorandum 1992-22
Juni 1992

vrije Universiteit *amsterdam*

The Riemannian interpretation of Gauss–Newton and Scoring, with application to system identification *

Ralf L.M. Peeters and Bernard Hanzon †‡

1 Introduction

The Gauss–Newton algorithm is an algorithm to obtain the solution(s) to nonlinear least squares problems. In system identification the algorithm is applied to e.g. the minimization of the sum of squares of the prediction errors, both for the off–line case and for the on–line case.
In [16] it was shown that the Gauss–Newton algorithm as used in prediction error algorithms for system identification can be interpreted approximately as a Riemannian gradient algorithm. In this paper this result is strengthened; it is shown that with a correct choice of the Riemannian metric, the Gauss–Newton algorithm can be interpreted as an *exact* Riemannian gradient algorithm. Furthermore this is now shown for general nonlinear least squares problems. This is especially interesting as the Gauss–Newton method is usually presented as an *approximation* to the Newton algorithm for this optimization problem. The general result is specialized to the case of prediction error algorithms for system identification and the corresponding Riemannian metrics are analysed. A central role is played by the so–called *prediction error metrics*. They are compared to the corresponding Fisher Information metrics and it is shown how they are related. As a result of this it follows that the Gauss–Newton algorithm is related (at least asymptotically) to the so–called method of scoring, which has in fact the interpretation of a Riemannian gradient algorithm if the metric is chosen to be the Fisher Information metric.
It is stressed that search algorithms which have the interpretation of Riemannian gradient algorithms are well–suited to be applied in so–called overlapping parametrization algorithms (see e.g. [38], [19], [39] and the references given there).
This paper is organized as follows. Sections 2 and 3 concern the Gauss–Newton algorithm, first in the Euclidean case, then on a differentiable manifold. In Sections 4 and 5 we present a framework on the basis of which the use of a Gauss–Newton prediction error algorithm for system identification can be motivated. Then, in Section 6, we discuss the closely related method of scoring, stemming from the field of statistics. We conclude with Section 7 where we present the results of a number of computer simulation experiments.

*This is an elaborate version of a paper to be presented at the IFAC Symposium MICC, Prague, September 1–2, 1992.

†Address: Free University, Faculty of Economics and Econometrics, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. E-mail: bhnz@sara.nl and ralf@sara.nl.

‡The research of the first author was carried out as part of NWO research project 611-304-019.

2 A new interpretation of the Gauss–Newton method for nonlinear least squares

One of the most often encountered approximation problems is the least squares problem. For the linear case the solution is very well known. Here we will focus on the nonlinear least squares problem. It usually arises if one tries to find approximate solutions to some system of equations by minimizing the sum of squares of the differences between left–hand side and right–hand side of the equations. These differences will be called the *errors* or *residuals*. Let $\mathcal{D} \subset \mathbf{R}^n$ be an open set. Consider a twice differentiable mapping [1] $f : \mathcal{D} \to \mathbf{R}^m$, $m \geq n$, $\theta \mapsto f(\theta)$ that will be called the *residual mapping*. Define the *least squares criterion* function

$$V(\theta) = \frac{1}{2}\|f(\theta)\|^2 = \frac{1}{2}f(\theta)^T f(\theta) = \frac{1}{2}\sum_{i=1}^{m} f^i(\theta)^2 \tag{2.1}$$

with f^i, $(i = 1, \ldots, m)$, denoting the components of f cq. the residuals, which are possibly nonlinear functions of the parameter vector $\theta = (\theta^1, \ldots, \theta^n)$. The nonlinear least squares problem is then to find the global minimum θ_* of the criterion function over the domain $\mathcal{D}$. For this problem usually iterative search procedures are employed. One of the disadvantages of this type of algorithm is that it will typically converge to a local minimum of the criterion function. If there happens to be only one local minimum, which is at the same time the global minimum and one can *show* this to be the case then the fact that the algorithm finds only local minima is not a problem. Otherwise one has to be more careful with the outcomes of the algorithm, but we will not go into that here. If convergence is assumed in the sequel, then this will be stated explicitly.
A general account of existing methods for function minimization can be found, e.g., in [6], [9]. The most well–known method to handle the nonlinear least squares problem is the method of Gauss–Newton. An excellent account of it can be found in [9, Ch. 10]. In this algorithm a number of iterations is performed. At each step a new estimate θ_+ is produced, given the current estimate θ_c, according to the formula

$$\theta_+ = \theta_c - [J(\theta_c)^T J(\theta_c)]^{-1} J(\theta_c)^T f(\theta_c) \tag{2.2}$$

Here, $J(\theta)$ denotes the Jacobian $\frac{\partial f}{\partial \theta}(\theta)$ of f at θ. By assumption it is continuous and even differentiable as a function of θ. The Gauss–Newton step (2.2) is well–defined only if $J(\theta_c)$ has full (column) rank, so that the inverse in (2.2) exists. Some remarks will be made about this in the next section. It will be assumed here that the Jacobian has full column rank, at all points encountered in the algorithm.
In literature one finds two standard views that motivate this algorithm. In the first one the algorithm is interpreted as an approximation to Newton's algorithm for the optimization problem at hand, in which a new estimate for θ_* is determined as

$$\theta_+ = \theta_c - H(\theta_c)^{-1} J(\theta_c)^T f(\theta_c), \tag{2.3}$$

with $H(\theta)$ denoting the Hessian of V at θ. Notice that $J(\theta)^T f(\theta)$ denotes the gradient of V at θ (as a column vector). The special structure of $V(\theta)$, being a sum of squares, leads to the following expression for the Hessian $H(\theta)$:

[1] If the mapping is only once continuously differentiable most of what follows remains valid. It is the mere possibility of comparing with methods that make use of second order derivatives, such as Newton's method, why we are making this assumption.

$$H(\theta) = J(\theta)^T J(\theta) + \sum_{i=1}^{m} f^i(\theta) \frac{\partial^2 f^i}{\partial \theta \partial \theta^T}(\theta) \tag{2.4}$$

If this Hessian is approximated by omitting the second term in the right–hand side of this expression one obtains the Gauss–Newton method, i.e. (2.3) becomes (2.2). One of the reasons that is given for omitting the second term is that it leads to an approximation of the Hessian that is positive definite (under the assumption of a full rank Jacobian). Therefore the inverse exists and moreover at each step the criterion function decreases, if necessary after modification of the step–size. Another argument is that one can obtain the first term from *first order* information only, which is usually much easier and cheaper to obtain than second order information. We notice that the approximation is expected to be good if the second term on the right–hand side is "small" compared to the first, which typically occurs in two situations: (a) if the residuals are (almost) linearly dependent on θ; (b) if the residuals are very small (something we are trying to achieve).
The second approach that leads to the Gauss–Newton algorithm is the so–called quasi–linearization approach. The idea is to use a linearization of f around the current estimate θ_c. The use of such a linearization is commonly called after Gauss, though it appears to be Legendre who originally introduced the concept (cf. [45]). Of course the linearization of f around the current estimate θ_c is given by

$$f^c(\theta) = f(\theta_c) + J(\theta_c)(\theta - \theta_c). \tag{2.5}$$

The next estimate is found by minimizing the approximate criterion function that corresponds to the linearization: $V^c(\theta) = \frac{1}{2}\|f^c(\theta)\|^2$. It is easily seen that θ_+ as calculated in the Gauss-Newton approach minimizes $V^c(\theta)$ and therefore the quasi–linearization approach leads again to the Gauss–Newton algorithm.
Here we propose a third interpretation of the Gauss-Newton algorithm, namely as a *Riemannian steepest descent algorithm* with a specific choice of the Riemannian metric. For the concepts of Riemannian geometry that are used here we refer to e.g. [7]. Let $\theta \in \mathcal{D}$. Assume again that the Jacobian $J(\theta)$ of the residual function f has full column rank. According to the inverse function theorem (cf. [7, pp.41–46]) there exists an open neighbourhood W of θ in $\mathcal{D}$ such that the mapping restricted to W, $f : W \to f(W)$ is a diffeomorphism with respect to the topologies and differentiable structures induced by $\mathbf{R}^n$ on W and by $\mathbf{R}^m$ on $f(W)$. In fact it follows that $f(W)$ is a (C^1-)differentiable submanifold of $\mathbf{R}^m$. The Euclidean metric on $\mathbf{R}^m$ therefore induces a Riemannian metric on the manifold $f(W)$ and using the diffeomorphism, it induces a Riemannian metric on W as well.

Proposition 2.1 *The tensor $\mathcal{R}$ of the Riemannian metric induced on W by the Euclidean metric on $\mathbf{R}^m$ via the mapping f can be expressed in local coordinates θ by the formula*

$$\|\dot{\theta}\|_{\mathcal{R}}^2 = \dot{\theta}^T J(\theta)^T J(\theta) \dot{\theta} \tag{2.6}$$

where $\|\cdot\|_{\mathcal{R}}$ denotes the associated norm on the tangent space $T_\theta(W)$ of W at the point θ and $\dot{\theta}$ denotes an element of $T_\theta(W)$ with respect to the natural basis induced by the local coordinates.

Proof When regarding a differentiable curve on $f(W)$, parametrized by $f(\theta(t))$ with $t \in (-\epsilon, \epsilon) \subset \mathbf{R}$ and such that $f(\theta(0)) = f(p)$ with tangent vector $\frac{d\theta}{dt}(0) = \dot{\theta}$, we see that the differential ds of the arclength at $f(p)$ is given by

$$ds = \|J(\theta)\dot{\theta}\|$$

On the other hand, this must be equal to

$$ds = \left(\dot{\theta}^T R(\theta)\dot{\theta}\right)^{1/2}$$

where $R(\theta)$ denotes the Riemannian metric tensor in terms of the local coordinates θ. Comparison of ds^2 for both expressions yields

$$R(\theta) = J(\theta)^T J(\theta) \tag{2.7}$$

since the above equations hold for all $\dot{\theta} \in \mathbf{R}^n$. This proves the proposition. □

Corollary 2.2 *If the open domain $\mathcal{D}$ of the mapping $f : \mathcal{D} \to \mathbf{R}^m$ consists of points at which the Jacobian of f has full column rank only, then the Euclidean metric on $\mathbf{R}^m$ induces a Riemannian metric on $\mathcal{D}$ via the mapping f. It is given in local coordinates by the formula in the previous proposition, i.e. $R(\theta) = J(\theta)^T J(\theta)$.*

Remark. Clearly $\mathcal{D}$ is a submanifold of $\mathbf{R}^n$, because it is an open subset. Together with the Riemannian metric as defined above, it is a Riemannian manifold. Locally $\mathcal{D}$ is a parametrization of the image space $f(\mathcal{D})$ in $\mathbf{R}^m$ and in fact f is locally an isometry of Riemannian manifolds. As for the global situation however, the set $f(\mathcal{D})$ does *not* have to be a submanifold of $\mathbf{R}^m$. In fact it does not have to be a manifold at all, let alone a Riemannian manifold.
For a Riemannian manifold M with local coordinates θ and Riemannian metric tensor expressed in these coordinates by $R(\theta)$ we have that the *Riemannian gradient* of a differentiable function V on M at a point p is given in the local coordinates by

$$R(\theta)^{-1}\frac{\partial V}{\partial \theta} \tag{2.8}$$

See for instance [1]. The Riemannian gradient describes, in local coordinates, the *maximizing normalized tangent vector* for V, where the normalization is with respect to the Riemannian metric at p. In other words, the Riemannian gradient describes the direction of steepest ascent of the criterion function with respect to the Riemannian metric. Using the Riemannian gradient one can define a Riemannian version of the method of steepest descent for the minimization of functions defined on M. Applying the Riemannian steepest descent algorithm to our Riemannian manifold $\mathcal{D}$ the following result is obtained.

Theorem 2.3 *Let $\mathcal{D}$ be an open domain of f, consisting of points for which the Jacobian of f has full column rank only, endowed with the Riemannian metric described in Corollary 2.2. Let f and V be defined as before. Consider a point $\theta \in \mathcal{D}$. Then the Riemannian steepest descent direction in θ coincides with the search direction of the Gauss–Newton method.*

Proof With respect to the local coordinates the ordinary gradient of $V(\theta) = \frac{1}{2}\|f(\theta)\|^2$ is given by $J(\theta)^T f(\theta)$. Using the general formula for the Riemannian gradient and substituting the formula for the Riemannian metric tensor found in Prop. 2.1, it follows that the Riemannian gradient is given by

$$[J(\theta)^T J(\theta)]^{-1} J(\theta)^T f(\theta)$$

Thus, the Riemannian steepest descent direction coincides in local coordinates θ with the Gauss–Newton direction, which proves the theorem. □

The geometrical interpretation that now results for the Gauss–Newton method is as follows. Consider the image set $f(\mathcal{D})$ in $\mathbf{R}^m$. The objective of minimizing $V(\theta)$ is equivalent to finding the point of $f(\mathcal{D})$ that is closest to the origin. Thus, being at the current iterate $\theta_c \in W \subset \mathcal{D}$, where W is an open neighbourhood of θ_c, as in Prop. 2.1, which corresponds to the point $f(\theta_c) \subset f(W)$ in the image space, a natural approach would be to calculate the orthogonal projection of the origin onto the tangent space to the manifold $f(W)$ at the point $f(\theta_c)$ (where the tangent space is regarded as an affine subset of $\mathbf{R}^m$). Using the vector from $f(\theta_c)$ pointing towards that optimum as the tangent vector determining the search direction to be explored, one obtains the Gauss–Newton search direction and by iteration the Gauss–Newton algorithm.

3 The Gauss-Newton algorithm acting on a differentiable manifold

If the domain of a function V to be minimized is a *differentiable manifold*, what is done in practice often comes down, from a geometrical point of view, to choosing a local coordinate chart in which an existing minimization algorithm can be applied, thereby implicitly using the *Euclidean* structure of the coordinate chart. As a consequence, starting from the same point on the manifold, one will generally obtain *different* iteration paths if different local coordinates are chosen. However, in case the manifold is endowed with a *Riemannian metric*, one can exploit this extra structure to obtain coordinate free algorithms for minimization over a manifold. Work in this direction can be found in [36], [32], [33], [11].
The geometrical interpretation of the Gauss–Newton algorithm presented in the previous section suggests that if the domain of f and of V is a *differentiable manifold*, a Gauss–Newton algorithm will be a straightforward generalization of the algorithm on an open subset of Euclidean space. The open subset of such a differentiable manifold of dimension n, say, that consists of all points at which the differential Df of f (which corresponds in local coordinates to the Jacobian of f) has full rank n is again a differentiable manifold. Let M denote this manifold. Locally it is diffeomorphic with an open subset of $\mathbf{R}^n$ and one can locally define a Riemannian metric, in the same way as in the previous section, as follows. Let $p \in M$ be a point on the manifold. Let W be an open neighbourhood of p in M such that it is diffeomorphic with an open subset of $\mathbf{R}^n$ and at the same time diffeomorphic with $f(W)$. Around each point of M such a neighbourhood exists according to the inverse function theorem. The Riemannian metric tensor $\mathcal{R}$ induced on $f(W)$ and W by the Euclidean metric on $\mathbf{R}^m$ can be expressed in local coordinates θ by the formula

$$\|\dot{\theta}\|_{\mathcal{R}}^2 = \dot{\theta}^T J(\theta)^T J(\theta)\dot{\theta} = \dot{\theta}^T R(\theta)\dot{\theta} \tag{3.1}$$

where $\|\cdot\|_{\mathcal{R}}$ denotes the associated norm on the tangent space $T_p(M)$ and $\dot{\theta}$ an element of $T_p(M)$ with respect to the natural basis induced by the local coordinates.
From the geometric interpretation of the Gauss–Newton algorithm it is immediately clear that the Riemannian metric thus obtained is *independent* of the local coordinates used on M and is also independent of any other Riemannian metric that one may or may not have defined on M. In this way the manifold M becomes a Riemannian manifold and at each point of the manifold the criterion function V will have a uniquely defined Riemannian steepest descent direction which is a vector in the tangent space which is *independent* of the choice of the local coordinates. The *representation* of that same tangent vector in terms of local coordinates θ is of course dependent on these local coordinates and is given by

$$-R(\theta)^{-1}\frac{\partial V}{\partial \theta} \tag{3.2}$$

Therefore the Gauss–Newton procedure is very well suited for nonlinear least squares problems for which the domain is a differentiable manifold!
Although the search direction at each point of the manifold is uniquely defined in a coordinate free fashion, the Gauss–Newton algorithm as it stands is *not* completely coordinate free. Indeed, the choice of local coordinates has some influence on the steps taken by the algorithm. The reason is that the recipe "take a step of a given length in a given direction" produces different points in different local coordinate charts. One way to make the algorithm completely coordinate–free is to make use of the *geodesics* on the manifold. Geodesics are well–defined on any Riemannian manifold and "take a step of a given length in a given direction along a geodesic" produces a uniquely defined point, independent of the local coordinates that are being used[2]. At this point it would appear to be most natural to use the geodesics derived from the Riemannian metric defined on the manifold M by the prescription that f be a local Riemannian isometry as before. However one might also use a completely different Riemannian metric on M in order to define the geodesics if on wishes to do so. For more details of the Gauss–Newton procedures on a differentiable manifold we refer to [39,40].
We conclude this section with three more remarks.
(i) For the points at which the Jacobian does not have full column rank n we notice that this may have two reasons. One is that an essential geometrical property is being lost (resulting in a drop of dimension of the tangent space). One cannot cure this. The other is that it is merely a result of a badly chosen parametrization. In case the image $f(\mathcal{D})$ is known to be an imbedded submanifold of dimension n in $\mathbf{R}^m$, then one can always reparametrize locally in order to obtain coordinates for which the Jacobian does not degenerate.
(ii) As the Gauss–Newton method in the current point of view is regarded as a Riemannian steepest descent method, the incorporation of a step–size controlling parameter α appears *naturally*, leading to the formula $\theta_+ = \theta_c - \alpha[J(\theta_c)^T J(\theta_c)]^{-1} J(\theta_c)^T f(\theta_c)$. This as opposed to the conventional approach, where such a parameter is always introduced as an *artificial* device to protect against step–sizes that might be too large. In those conventional philosophies $\alpha = 1$ would be optimal, as it optimizes the quasi–linearized criterion function.
(iii) Suppose that the Df has full column rank n at the true optimum $p_* \in M$. Then, for the sake of convergence analysis of the Gauss–Newton algorithm, it is an interesting result that there exists an open neighbourhood W of p_* on which Df has full column rank and which can be parametrized by local coordinates θ, with the property that the Gauss–Newton algorithm remains in it, provided the mechanism by which steps are taken is such that a decrease of function value is guaranteed. This is e.g. the case if geodesics or lines are followed on which one tries to find a minimum of V, the so–called *line search*. For a proof of this statement, see Appendix A.

4 Manifolds of prediction error filters

In the following sections we shall describe a theoretical framework on the basis of which we can motivate the use of a Gauss–Newton prediction error algorithm for system identification. The work of the present section is based on [16,17].
Let us consider a p–dimensional stationary Gaussian process $Y_{-\infty}^{\infty} = \{y_t\}_{t=-\infty}^{\infty}$ with rational

[2]This holds under the condition that the step–length is not too large, otherwise one might be thrown off the manifold.

spectrum having $2N$ poles, multiplicities included, and which has no zeroes on the unit circle. It is a standard result from stochastic realization theory that this process can be modelled by a state space model of the following form (the *innovations representation*):

$$S(\theta_*) \; : \; \left\{ \begin{array}{rcl} x_{t+1} & = & A(\theta_*)x_t + B(\theta_*)v_t \\ y_t & = & C(\theta_*)x_t + v_t \end{array} \right. \tag{4.1}$$

with $\{v_t\}_{t=-\infty}^{\infty}$ a p–dimensional Gaussian white noise process of zero mean and covariance $\Sigma > 0$: $v_t \sim N(0, \Sigma)$; with both $A(\theta_*)$ and $A(\theta_*) - B(\theta_*)C(\theta_*)$ asymptotically stable matrices of size $N \times N$; and with the triple $(A(\theta_*), B(\theta_*), C(\theta_*))$ a minimal realization, that is, $(A(\theta_*), B(\theta_*))$ is controllable and $(C(\theta_*), A(\theta_*))$ is observable. Thus, the dimension of the state space is N.

In order to allow for a simplification in the *interpretation* of the Gauss–Newton algorithm that is to follow, we will consider the stochastic processes from $t = 1$ onwards and make the stylized assumption that the initial state x_1 is *known to be equal to zero.* This assumption is *not* required for the *construction* of the Gauss–Newton procedures.

It is well–known that the minimal state space representation of the input/output system $S(\theta_*)$ is unique up to a choice of state space basis. To obtain local identifiability, it will be assumed that $(A(\theta_*), B(\theta_*), C(\theta_*))$ is put in some suitable local canonical form. To describe the set of all i/o–systems that are relevant for the identification procedure use will be made of local coordinates θ in some local coordinate chart Θ which is an open subset of $\mathbf{R}^n$. As in Section 3 the results can be extended to the manifold case in a rather straightforward manner, but here we will restrict ourselves to the case of *one* coordinate chart. The matrices $A(\theta), B(\theta)$ and $C(\theta)$ are assumed to depend differentiably on the local coordinates θ and to be in the same local canonical form as $(A(\theta_*), B(\theta_*), C(\theta_*))$. For examples of (overlapping!) local canonical forms we refer to [38], [21]. In order to extract information from Y_1^{∞} about parameter vector θ_* we can apply linear filtering to it. We define $\mathcal{H}$ to be the space of all linear mappings $h : Y_1^{\infty} \to \mathbf{R}^p$ such that h has finite covariance. Obviously we can associate with each $h \in \mathcal{H}$ a unique sequence of $p \times p$ matrices $\{H_1, H_2, H_3, \ldots\}$ such that

$$h = H_1 y_1 + H_2 y_2 + H_3 y_3 + \cdots \tag{4.2}$$

In this context, the requirement for h to have finite covariance comes down to the requirement $\mathrm{tr} \sum_{k=1}^{\infty} H_k H_k^T < \infty$. We can make $\mathcal{H}$ into a Hilbert space by introducing the inner product $\langle \cdot, \cdot \rangle$ as

$$\forall h, \tilde{h} \in \mathcal{H} \; : \quad \langle h, \tilde{h} \rangle = \mathbf{E}_{\theta_*} h^T \tilde{h}. \tag{4.3}$$

Here $\mathbf{E}_{\theta_*}$ denotes expectation with respect to the true underlying probability measure. It is easily verified that the above indeed constitutes a well–defined inner product on $\mathcal{H}$.

Next, we can consider the set of prediction error filters of order N, given by the recurrence equations

$$\Phi(\theta) \; : \; \left\{ \begin{array}{rcl} \hat{x}_{t+1} & = & [A(\theta) - B(\theta)C(\theta)]\hat{x}_t + B(\theta)y_t \\ \epsilon_t & = & -C(\theta)\hat{x}_t + y_t \end{array} \right. \tag{4.4}$$

with fixed initial state $\hat{x}_1 = 0$ and t ranging over $\mathbf{Z}^+$.

Then for each value of t the filters $\Phi(\theta)$ define a mapping $\epsilon_t : \Theta \to \mathcal{H}, \theta \mapsto \epsilon_t(\theta)$. Indeed, the asymptotic stability requirement with respect to $A(\theta)$ and $A(\theta) - B(\theta)C(\theta)$ establishes the finite variance property of $\epsilon_t(\theta)$ and in fact the limit of this variance for $t \to \infty$ exists and is finite. For a *fixed* value of t, consider the image set $\epsilon_t(\Theta) := \{\epsilon_t(\theta) \,|\, \theta \in \Theta\}$ of the

mapping ϵ_t. Assume t to be sufficiently large. Then this set forms a submanifold of the Hilbert space $\mathcal{H}$. (For a proof of an analogous result see [18].) The inner product on $\mathcal{H}$ induces via the mapping ϵ_t a Riemannian metric on Θ with Riemannian metric tensor given in local coordinates $\theta \in \Theta$ by

$$R_t(\theta) = \mathbf{E}_{\theta_*}(\frac{\partial \epsilon_t}{\partial \theta})^T(\frac{\partial \epsilon_t}{\partial \theta}). \tag{4.5}$$

Here, one should remark that for each coordinate θ^i of θ we have that also the mapping $\frac{\partial \epsilon_t}{\partial \theta^i}$ is an element of $\mathcal{H}$, as follows from the fact that this derivative mapping can also be obtained via a (somewhat larger) linear filter, which is again i/o–stable. The metric on Θ that is obtained this way will be called a prediction error metric. Its limit for $t \to \infty$ exists and the formula for the Riemannian metric tensor, denoted here by $R_\infty(\theta)$, is obtained by substituting in (4.5) the filters that arise if one starts the process and the filters at $-\infty$. This is the metric obtained in [16].

5 Geometrical interpretation of the Gauss–Newton algorithm for system identification

A well–known criterion function $V_t(\theta)$ for system identification from a data set of t observations, is half the mean square of the norms of the prediction error vectors that are obtained if the parameter estimate is θ:

$$\hat{V}_t(\theta) = \frac{1}{2t}\sum_{k=1}^{t} \epsilon_k(\theta)^T \epsilon_k(\theta) \tag{5.1}$$

For the background of this criterion see e.g. [46]. Clearly in this case the system identification problem is a nonlinear least squares problem and the results of the earlier sections are applicable. It follows that if the $n \times n$ matrix given by

$$\hat{R}_t(\theta) = \frac{1}{t}\sum_{k=1}^{t}(\frac{\partial \epsilon_k(\theta)}{\partial \theta})^T(\frac{\partial \epsilon_k(\theta)}{\partial \theta}). \tag{5.2}$$

has full rank n, then it is the Riemannian metric tensor which makes the mapping $\Theta \to \mathbf{R}^{pt}, \theta \mapsto (\epsilon_1(\theta)^T, \ldots, \epsilon_t(\theta)^T)^T$ into a Riemannian isometry as before.

Proposition 5.1 *With probability one it holds that* $\lim_{t\to\infty} \hat{R}_t(\theta) = R_\infty(\theta)$ *the Riemannian metric tensor of the steady state prediction error metric.*

Proof The proposition follows from the ergodicity and stability properties of the stochastic processes involved. Indeed, as shown, e.g., in [19, p.232] one can obtain the derivatives of the prediction errors via linear filtering of the data with an asymptotically stable linear filter, obtained by adding partial derivatives of the equations determining $\Phi(\theta)$ to the equations of $\Phi(\theta)$ itself. □

Corollary 5.2 *Let θ be given. With probability one, for t large enough $\hat{R}_t(\theta)$ is positive definite.*

Proof This is a consequence of the fact that R_∞ denotes a Riemannian metric tensor, and therefore is positive definite. See [16]. □

It is likely that with some more work one can show that for a certain value of t with

probability one $\hat{R}_t(\theta)$ is positive definite for all $\theta \in \Theta$. In any case one can say that for all θ for which the Gauss–Newton direction is defined it coincides with the Riemannian steepest descent direction, which is in this case given by $-\left(\hat{R}_t(\theta)\right)^{-1} \frac{\partial \hat{V}_t(\theta)}{\partial \theta}$. Just as in the general case of nonlinear least squares it follows that the Gauss–Newton direction is independent of the choice of the local coordinates and therefore is very well suited for an overlapping parametrizations approach.
Consider the following theoretical criterion function for each value of $t \in \mathbf{Z}^+$:

$$V_t(\theta) = \frac{1}{2}\mathbf{E}_{\theta_*}\{\epsilon_t(\theta)^T \epsilon_t(\theta)\} \tag{5.3}$$

One easily shows that θ_* constitutes a global minimum for each $V_t(\theta)$ as a consequence of the assumption $x_1 = 0$. If t is large enough this minimum is *unique* (as a global one) as we show in Appendix A. The limit of $V_t(\theta)$ for $t \to \infty$ exists and is denoted by $V_\infty(\theta)$. It is equal to half the expected value of the norm squared of the steady state prediction errors, i.e. that are obtained if one assumes that the process and the filters have started running at $-\infty$. For large values of t the actual criterion function $\hat{V}_t(\theta)$ will approach $V_\infty(\theta)$. So for large values of t the Gauss–Newton algorithm will behave approximately like the Riemannian steepest descent algorithm for the *theoretical* criterion function $V_\infty(\theta)$, with respect to the Riemannian metric induced by the steady state prediction error metric, i.e. the Riemannian metric given in local coordinates by $R_\infty(\theta)$. Both $V_\infty(\theta)$ and $R_\infty(\theta)$ depend on the true value θ_* as well as on θ and therefore cannot be calculated in a system identification algorithm, as θ_* is of course unknown.
As is well–known, also $V_\infty(\theta)$ has a unique global maximum at the true value θ_* (see e.g. [46]). Therefore the only obstacle to consistency of the procedure is the possible existence of local, non-global minima. This is of course a problem for all search algorithms of this sort. For the case of on–line identification the Gauss-Newton procedure as constructed in e.g. [35] has asymptotically the interpretation of a Riemannian gradient algorithm, *provided that* a weighting is applied that makes the matrix that is supposed to approximate the Riemannian metric tensor flexible enough with respect to changes in the parameter. In fact this condition may be a rather strong one (see [16]). Therefore it may be advisable in the case of on–line identification to use a different, but related method, namely the method of scoring, which will be treated in the next section.

6 A Riemannian interpretation of the scoring algorithm and its relation with the Gauss–Newton algorithm for system identification

In this section we discuss the so–called *method of scoring*, as described in [43, pp.366–374]. As will be shown, this method provides an alternative to the method of Gauss–Newton that can be viewed as a Riemannian gradient method acting on the system manifold as well.
In literature there appears to be a widespread confusion about the exact definition of the method of scoring. Taking [43] as our point of departure, we shall point out the differences between various definitions and discuss how some occurring methods are interrelated.
The scoring method stems from the field of statistics and is closely related to the method of *maximum likelihood*. As such, its applicability extends beyond that of the present paper, where the likelihood function $\hat{L}_t(\theta)$ for the prediction errors $\{\epsilon_k(\theta)\}_{k=1}^t$ exhibits a special structure. The so–called method of *linearized maximum likelihood*, cf. [47, p.527], proceeds by determining a new estimate θ_+ for θ_* according to

$$\theta_+ = \theta_c - \hat{H}_t(\theta_c)^{-1}\hat{g}_t(\theta_c) \tag{6.1}$$

where $\hat{g}_t(\theta_c)$ and $\hat{H}_t(\theta_c)$ denote the (ordinary) gradient and Hessian, respectively, of the average negative log–likelihood $-\frac{1}{t}\log\hat{L}_t(\theta)$ at θ_c. This can be regarded as a step taken by *Newton's method*, and it is *not* equivalent to scoring as we shall see.
To obtain the scoring algorithm, we replace the Hessian $\hat{H}_t(\theta_c)$ at θ_c by its expectation based on the probability measure stemming from θ_c: we take the expectation as if θ_c were the true underlying parameter vector. This gives the *Fisher information* at θ_c instead, as is proved, e.g., in [46, App. B.4]. Denoting the average Fisher information $t\mathbf{E}_{\theta_c}\hat{g}_t(\theta_c)\hat{g}_t(\theta_c)^T$ at θ_c by $I_t(\theta_c)$ we arrive at the scheme

$$\theta_+ = \theta_c - I_t(\theta_c)^{-1}\hat{g}_t(\theta_c) \tag{6.2}$$

A clear definition of scoring in accordance with this formula and an exposition of its relation to Gauss–Newton can be found in [22, p.131–135].
As opposed to both Gauss–Newton and linearized maximum likelihood, the matrix $I_t(\theta_c)^{-1}$ premultiplying the gradient of $-\frac{1}{t}\log\hat{L}_t(\theta)$ at θ_c does, apart from its dependence on the current estimate θ_c, *not* depend on the measured data. This is indicated by omission of the caret. Scoring can be interpreted as a *Riemannian steepest descent method* on a manifold of probability densities. This is a consequence of the following basic theorem.

Theorem 6.1 *Suppose the set of probability densities that one wants to consider in an estimation procedure forms a differentiable manifold such that at all points on the manifold the Fisher information matrix is well–defined and positive definite. Then the Fisher information matrix has the interpretation of a Riemannian metric tensor. It defines a Riemannian metric, the so–called Fisher metric, on the manifold of densities.*

Proof This statement can be found in rudimentary form in [43, p.332]. For a proof see e.g. [5], [2]. □

As a consequence of this theorem, scoring can be made completely *coordinate free*, at least in principle, by taking steps along geodesics in the proposed directions. This is entirely analogous to the procedure of obtaining coordinate free versions of Gauss–Newton.
Now let us apply this to the problem of system identification as posed in the previous sections. To make the connection we will assume that the covariance matrix of the innovations is known to be the identity matrix. We will consider the situation for large values of t, such that effects of initial conditions can be neglected. From e.g. [46, Sect. 7.4] it follows that in the Gaussian set–up of the identification problem studied in this paper the *negative log–likelihood* is given by

$$-\log\hat{L}_t(\theta) = \frac{1}{2}\sum_{k=1}^{t}\epsilon_k(\theta)^T\epsilon_k(\theta) + \text{constant} \tag{6.3}$$

where, as stated above, the effects of initial conditions are neglected and the covariance of the driving white noise is assumed to be $\Sigma = I$. Now, maximization of the likelihood function is equivalent to minimization of $\hat{V}_t(\theta)$ as defined previously, since we can write $-\frac{1}{t}\log\hat{L}_t(\theta) = \hat{V}_t(\theta) + \text{constant}$. Therefore, Gauss–Newton can be applied. We get

$$\theta_+ = \theta_c - \hat{G}_t(\theta_c)^{-1}\hat{g}_t(\theta_c) \tag{6.4}$$

This Gauss–Newton method is clearly *not* identical to scoring, as opposed to what is stated in [34, p.284].

Consider for the system identification case the limit for $t \to \infty$ of the average Fisher information: $I_\infty(\theta) = \lim_{t\to\infty} I_t(\theta)$. One has a result similar to the previous one. In this case one considers the situation in which one wants to identify a stochastic system. As before we will speak of an i/o–system if we want to stress that only the external behaviour of the system is considered. In this case each i/o–system is in fact a stationary stochastic process with rational spectrum.

Theorem 6.2 *Suppose the set of i/o–systems that one wants to consider in an estimation procedure forms a differentiable manifold such that at all points on the manifold the Fisher information matrix is well–defined and positive definite. Then the Fisher information matrix has the interpretation of a Riemannian metric tensor. It defines a Riemannian metric, the so–called Fisher metric, on the manifold of i/o–systems.*

Proof See [27], [3]. □

In order to compare Gauss–Newton and scoring for our system identification problem, we can compare the metrics involved, as both methods have the interpretation of a Riemannian steepest descent method. For large values of t the prediction error metric, which is associated to Gauss–Newton, approaches the steady state prediction error metric. Therefore we will compare the steady state prediction error metric with the Fisher metric. Calculation of the elements of the corresponding metric tensor proceeds in terms of the matrices $A(\theta)$, $B(\theta)$, $C(\theta)$ via solving discrete–time Lyapunov equations associated with the extended system

$$\begin{cases} \zeta_{t+1} &= \tilde{A}(\theta_*,\theta)\zeta_t + \tilde{B}(\theta_*,\theta)v_t \\ \dot{\epsilon}_t &= \tilde{C}(\theta_*,\theta)\zeta_t \end{cases} \tag{6.5}$$

Here $\dot{\epsilon}_t$ denotes the directional derivative of ϵ_t in the direction $\dot{\theta}$, whereas $\tilde{A}(\theta_*,\theta)$, $\tilde{B}(\theta_*,\theta)$, $\tilde{C}(\theta_*,\theta)$ are given by

$$\begin{aligned} \tilde{A}(\theta_*,\theta) &= \begin{pmatrix} A(\theta_*) & 0 & 0 \\ B(\theta)C(\theta_*) & F(\theta) & 0 \\ \dot{B}(\theta)C(\theta_*) & \dot{F}(\theta) & F(\theta) \end{pmatrix} \\ \tilde{B}(\theta_*,\theta) &= \begin{pmatrix} B(\theta_*) \\ B(\theta) \\ \dot{B}(\theta) \end{pmatrix} \\ \tilde{C}(\theta_*,\theta) &= \begin{pmatrix} 0 & -\dot{C}(\theta) & -C(\theta) \end{pmatrix}, \end{aligned} \tag{6.6}$$

with $F(\theta) = A(\theta) - B(\theta)C(\theta)$ and the dot denoting directional differentiation. Thus, the prediction error metric can be related to the ℓ^2–norm of systems of order $\leq 3N$. The Fisher information matrix is obtained from the *same formulas* via substitution of $\theta_* = \theta$. It is an interesting fact that the triple $(\tilde{A}(\theta,\theta), \tilde{B}(\theta,\theta), \tilde{C}(\theta,\theta))$ is *non–minimal* and can be shown to correspond to a system of order $\leq 2N$. For more details we refer to [39].
A number of things can be said if the actual estimate θ_c approaches the true value θ_* and t tends to ∞. In that case the Hessian of the likelihood is known to converge to the information matrix and, as we have seen, this in turn becomes equal to the steady state prediction error metric tensor. Therefore, if t is sufficiently large and θ_c close enough to θ_* then the exact Newton method, Gauss–Newton and the scoring algorithm all produce virtually the same search directions at θ_c so that their convergence behaviour is comparable. Of these three methods the Newton method is the only one that is *not* coordinate–free,

because in general the Hessian of a function is not a coordinate free object. There is however one exception: at critical points of a function the Hessian *is* coordinate–free. Of course in the situation just described one has a critical point of the criterion function indeed and the Hessian does become a coordinate–free object.

7 Simulation experiments

We have carried out several computer experiments based on the off–line identification set–up described in the previous sections. Using a simulated data sample of size 2000 for a specific system of order $N = 4$ and with $p = 2$ inputs and outputs, we have estimated its $n = 2Np = 16$ system parameters in various different ways. Some results on required numbers of iterations are collected in Tables 1–3. The simulated data correspond to *steady state* simulations and do *not* involve the stylized assumption $x_1 = 0$ of Section 4 for the data generating process. Of course the prediction error filters did start with initial state zero. In order to detect numerical convergence within a given parametrization we made use of Marquardt's stopping criterion as described in [6].
To handle the problem of selecting an appropriate parametrization we have used the *structure selection criterion* described in [37,38]. According to that approach, for $N = 4$ and $p = 2$ there exist three different parametrizations, based on different nice selections. These correspond to matrices A, B and C, parametrized respectively by

- Structure 1:

$$A = \begin{pmatrix} 0 & 0 & 1 & 0 \\ \theta_1 & \theta_2 & \theta_3 & \theta_4 \\ 0 & 0 & 0 & 1 \\ \theta_5 & \theta_6 & \theta_7 & \theta_8 \end{pmatrix} \qquad B = \begin{pmatrix} \theta_9 & \theta_{13} \\ \theta_{10} & \theta_{14} \\ \theta_{11} & \theta_{15} \\ \theta_{12} & \theta_{16} \end{pmatrix}$$

- Structure 2:

$$A = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \theta_1 & \theta_2 & \theta_3 & \theta_4 \\ \theta_5 & \theta_6 & \theta_7 & \theta_8 \end{pmatrix} \qquad B = \begin{pmatrix} \theta_9 & \theta_{13} \\ \theta_{10} & \theta_{14} \\ \theta_{11} & \theta_{15} \\ \theta_{12} & \theta_{16} \end{pmatrix}$$

- Structure 3:

$$A = \begin{pmatrix} \theta_1 & \theta_2 & \theta_3 & \theta_4 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \theta_5 & \theta_6 & \theta_7 & \theta_8 \end{pmatrix} \qquad B = \begin{pmatrix} \theta_9 & \theta_{13} \\ \theta_{10} & \theta_{14} \\ \theta_{11} & \theta_{15} \\ \theta_{12} & \theta_{16} \end{pmatrix}$$

Matrix C assumes in all three cases the fixed form

$$C = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

To generate the data sample we have constructed the driving white noise as Gaussian, 2–dimensional, with zero mean and *unit covariance*, that is

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

The data generating system of our experiments is then characterized by the "true" parameter vector

$$\theta_* = (0.1, -0.3, 0.1, 0.1, -0.3, -0.7, 0.1, -0.1, 0.1, 0.2, -0.2, -0.1, -0.1, 0.1, 0.1, 0.1)^T$$

corresponding to structure 2.
We have investigated the Gauss–Newton and scoring algorithm, both starting from the same 4 different initial points. These starting points are given by

$$\theta_0^{(1)} = (0.2, -0.4, 0.0, 0.2, -0.4, -0.8, 0.2, 0.0, 0.0, 0.3, -0.1, -0.2, -0.2, 0.2, 0.2, 0.0)^T$$

$$\theta_0^{(2)} = (0.0, 0.1, 0.1, -0.1, -0.2, 0.1, 0.0, -0.1, 0.1, -0.2, 0.1, -0.1, 0.0, -0.1, 0.1, -0.2)^T$$

$$\theta_0^{(3)} = (0.4, -0.2, 0.3, -0.1, -0.2, -0.1, -0.1, -0.1, 0.2, -0.2, 0.1, 0.1, -0.3, -0.2, 0.3, 0.1)^T$$

$$\theta_0^{(4)} = (0.5, 0.0, 0.1, -0.1, -0.2, 0.1, -0.3, -0.1, 0.1, 0.2, -0.3, -0.1, 0.0, -0.2, 0.0, -0.2)^T$$

where $\theta_0^{(1)}$, $\theta_0^{(3)}$ and $\theta_0^{(4)}$ are in structure 2, and $\theta_0^{(2)}$ is in structure 1.
The required numbers of iterations needed to converge to a local optimum are collected in Table 1. About the different starting points we remark that $\theta_0^{(1)}$ is closest to θ_*. This is well reflected by the results of the experiments, as convergence was always quickest starting from this point.
A second series of experiments corresponds to the same two algorithms, both starting from the same four initial points but this time with the initial parameter chart kept fixed. The outcomes of these experiments are collected in Table 2. As the first–order effects of the algorithms are parametrization independent, the differences can be assigned fully to "second–order effects."
The third series of experiments involves starting point $\theta_0^{(2)}$ which is analyzed for all three parametrizations, which are kept fixed each time. For this purpose, starting point $\theta_0^{(2)}$ is recalculated to its equivalents in structure 2 and 3. The results are collected in Table 3.
Based on these experiments we then can draw the following conclusions.

1. The use of overlapping parametrizations can prove to be essential for finding the true optimum in a system identification problem. In those cases where the wrong *fixed* parameter chart was used, both Gauss–Newton and scoring were not able to approximate the true optimum satisfactory. This is reflected by the results of Table 2 and 3. However, when automatic changing of parametrization was applied no such problems occurred and the true optimum was approximated well by Gauss–Newton and scoring. As already remarked above, these effects can be fully assigned to the fact that we did not follow geodesics on the manifold, but that steps were taken along lines within the parameter charts instead. Thus, the manifold approach with overlapping parametrizations may improve existing identification methods. In conjunction with this, notice that also if one starts in the correct structure convergence might be much slower if this parametrization is kept fixed. See Table 3.

2. Local convergence in the neighbourhood of the true optimum for Gauss–Newton and scoring occurred at similar rates, which was superlinear.

We conclude the discussion of these experiments with some final remarks.
(i) The lack of built–in "hysteresis" in the structure selection algorithm (as proposed by Clark [8]) caused no problems in our experiments. Such a facility seems only necessary

method	starting point			
	$\theta_0^{(1)}$	$\theta_0^{(2)}$	$\theta_0^{(3)}$	$\theta_0^{(4)}$
Gauss–Newton	26	48	33	34
Scoring	29	33	34	30

Table 1. Number of iterations required for convergence in case of automatic structure switching.

method	starting point			
	$\theta_0^{(1)}$	$\theta_0^{(2)}$	$\theta_0^{(3)}$	$\theta_0^{(4)}$
Gauss–Newton	26	> 100	33	34
Scoring	29	> 100	33	30

Table 2. Number of iterations required for convergence in case of fixed parametrizations.

if the optimum corresponds to a model that lies close to the boundary of an area in the admissable space related to a fixed optimal structure. But even then the algorithms can terminate appropriately as the number of possible charts is finite. In our set–up it is immaterial to which structure the final estimated model corresponds.
(ii) From some additional experimenting we have found that scoring is a more robust method than Gauss–Newton, because of the following. If one starts relatively far from the true optimum, one might be "thrown off the manifold." This is caused by the fact that the criterion function to be optimized remains well defined as long as $A(\theta) - B(\theta)C(\theta)$ is asymptotically stable, without any restriction on the asymptotic stability of $A(\theta)$. Thus, in principle it can happen for any optimization method that a path is followed that leads through the region where $A(\theta)$ is unstable. As a consequence, local minima on that stability boundary occur, a situation that cannot be avoided. However, one should notice that the Riemannian metric tensor related to the method of scoring, i.e. the Fisher information matrix, "explodes" if θ approaches the stability boundary for either $A(\theta)$ or $A(\theta)-B(\theta)C(\theta)$. Therefore, this method implicitly tries to avoid the situation mentioned above, as opposed to Gauss–Newton which does not exhibit this behaviour.

method	parametrization		
	1	2	3
Gauss–Newton	> 100	41	> 100
Scoring	> 100	> 100	> 100

Table 3. Number of iterations required for convergence in case of fixed parametrizations, using starting point $x_0^{(2)}$.

Appendix A :

We start this appendix by showing that the open neighbourhood W of p_* in M can be taken of a special form around an isolated local minimum of the criterion function at which the Jacobian has full column rank n. As stated in Section 3, this is important in the light of convergence analysis of the algorithms. We have the following.

Lemma A.1 *Suppose $f : \mathcal{D} \to \mathbf{R}^m$ is a continuously differentiable mapping, where $\mathcal{D} \subset \mathbf{R}^n$ is an open domain and $m \geq n$. Denote its Jacobian by $J : \mathcal{D} \to \mathbf{R}^{m \times n}$ defined as $J(\theta) = \left(\frac{\partial f}{\partial \theta^1}(\theta), \ldots, \frac{\partial f}{\partial \theta^n}(\theta)\right)$, $\forall \theta \in \mathcal{D}$. Define function $V : \mathcal{D} \to \mathbf{R}$ as $V(\theta) = \frac{1}{2} \parallel f(\theta) \parallel^2$, with $\| \cdot \|$ denoting the Euclidean norm on $\mathbf{R}^m$. Suppose that θ_* yields a local, isolated minimum of V on D, for which the Jacobian $J(\theta_*)$ has full rank n. For all $\delta \geq 0$ define $D^*_\delta \subset D$ as the open set*

$$D^*_\delta = \{\theta \in D \,|\, V(\theta) - V(\theta_*) < \delta;\ \theta \text{ and } \theta_* \text{ are arcwise connected } \} \tag{A.1}$$

Then there exists $\delta > 0$ for which $rk\{J(\theta)\} = n$, $\forall \theta \in D^*_\delta$.

Proof According to the inverse function theorem there exists an open neighbourhood W of θ_* such that for all $\theta \in W$: $\mathrm{rk}\{J(\theta)\} = n$, and for which the restriction $f|_W : W \to f(W)$ of f to W with range $f(W)$ is bijective.
Since W is open in $\mathbf{R}^n$ there exists an $\epsilon > 0$ for which the open ball $B_\epsilon(\theta_*) = \{\theta \in \mathbf{R}^n \mid \| \theta - \theta_* \| < \epsilon\}$ is contained in W. (Here $\| \cdot \|$ again denotes the Euclidean norm, but this time on $\mathbf{R}^n$.)
Since θ_* yields an isolated local minimum of V and because V is continuous on D, there exists an $\eta > 0$ such that for all $\theta \in B_\eta(\theta_*)$ we have $V(\theta) > V(\theta_*)$ if $\theta \neq \theta_*$.
Denote $\mu = \frac{1}{3}\min(\epsilon, \eta)$ and consider the closed annulus $A^*_{\mu,2\mu} = \{\theta \in \mathbf{R}^n \,|\, \mu \leq \| \theta - \theta_* \| \leq 2\mu\}$. This is a compact subset of D so that the restriction of V to $A^*_{\mu,2\mu}$ assumes a minimum value, say $\tilde{V}(\mu, 2\mu)$. Define $\Delta = \tilde{V}(\mu, 2\mu) - V(\theta_*)$. Since $A^*_{\mu,2\mu}$ is contained in $B_\eta(\theta_*)$ but does not contain θ_* we must have $\Delta > 0$.
Now consider the set D^*_Δ. Obviously we must have $D^*_\Delta \subset B_\mu(\theta_*)$, since all $\theta \in D^*_\Delta$ are by definition arcwise connected to θ_* but no arc in D^*_Δ can have points in $A^*_{\mu,2\mu}$: on $A^*_{\mu,2\mu}$ we have $V(\theta) - V(\theta_*) \geq \Delta$ but on D^*_Δ we have by definition $V(\theta) - V(\theta_*) < \Delta$.
We can summarize: $D^*_\Delta \subset B_\mu(\theta_*) \subset B_\epsilon(\theta_*) \subset W$, and since for all $\theta \in W$ the Jacobian of f has full rank n, this also holds for all $\theta \in D^*_\Delta$. Thus, any $\delta \in (0, \Delta]$ will do, which proves the proposition. □

Corollary A.2 *From Lemma A.1 it follows that if the step-size controlling parameter α in the Gauss-Newton scheme is determined according to a rule that ensures a decrease of function value, then around each local minimum that is isolated and for which $J(\theta)$ has full column rank n there exists a neighbourhood such that if the algorithm enters it, it can never leave it. Within this region there is no more need of changing coordinates, so that the convergence theorems developed for the Euclidean case can readily be applied.*

Remark. Obviously, if the domain for f and V is not a subset of $\mathbf{R}^n$ but an n-dimensional Riemannian manifold, we can first restrict to local coordinates in a chart containing p_* and then apply Lemma A.1 above. This shows that our claims are true for the case of Riemannian manifolds as well. Apart from that, one can state corollaries like the one above for *any* minimization algorithm that is such that the sequence of function values for the

iterates is monotonously decreasing.

The following discussion corresponds to some claims in Section 5.
Let $(A(\theta), B(\theta), C(\theta))$ be a parametrized matrix triple, where $\theta \in \Theta \subset \mathbf{R}^d$, with Θ an open set, such that $A : \Theta \to \mathbf{R}^{n\times n}$, $B : \Theta \to \mathbf{R}^{n\times p}$ and $C : \Theta \to \mathbf{R}^{p\times n}$ are sufficiently often continuously differentiable functions of θ. We assume the functions A, B and C and the set Θ to be such that the following properties are satisfied: (1) $(A(\theta), B(\theta), C(\theta))$ is minimal for all $\theta \in \Theta$, (2) $A(\theta)$ and $[A(\theta) - B(\theta)C(\theta)]$ are asymptotically stable, (3) $(A(\theta), B(\theta), C(\theta))$ is the only representative of its i/o–equivalence class in Θ.
It is known from realization theory that, in order for these conditions to be satisfied, the dimension d of the parameter space can be at most $2np$.
To each triple $(A(\theta), B(\theta), C(\theta))$ we can associate a linear, time–invariant, discrete–time, dynamical system, denoted by $S(\theta)$, according to the following system of recursive equations:

$$S(\theta): \quad \begin{cases} x_{t+1}(\theta) = A(\theta)x_t(\theta) + B(\theta)v_t \\ y_t(\theta) = C(\theta)x_t(\theta) + v_t \end{cases} \qquad (t \in \mathbf{Z}^+), \quad x_1(\theta) = 0,$$

with $\{v_t\}_{t=1}^{\infty}$ a p–dimensional, stationary, white, Gaussian noise process of zero mean and with covariance $\Sigma > 0$.
Notice that in this set–up the initial state of the system is known and required to be zero. It is the price we have to pay in order to be able to carry out the following analysis and to obtain the results collected in this Appendix.
The system above is given in so–called innovations form; the v_t are the innovations. For any Σ (i.e., for any input process $\{v_t\}_{t=1}^{\infty}$) the system $S(\theta)$ determines an output process denoted by $Y_1^{\infty}(\theta) = \{y_t(\theta)\}_{t=1}^{\infty}$ which is zero mean and Gaussian, but in general not white.
Essentially we consider dynamical systems to be i/o–mappings. Our interest is in extracting information about the parameter θ (within a given parametrization) from the output process $Y_1^{\infty}(\theta)$. For this we consider linear filtering of an output process $Y_1^{\infty} = \{y_t\}_{t=1}^{\infty}$ by filters $\Phi(\theta)$ of the form

$$\Phi(\theta): \quad \begin{cases} \hat{x}_{t+1}(\theta) = [A(\theta) - B(\theta)C(\theta)]\hat{x}_t(\theta) + B(\theta)y_t \\ \hat{y}_t(\theta) = C(\theta)\hat{x}_t(\theta) \end{cases} \qquad (t \in \mathbf{Z}^+), \quad \hat{x}_1(\theta) = 0,$$

with $\theta \in \Theta$.
Introducing the prediction error process $E_1^{\infty}(\theta) = \{\epsilon_t(\theta)\}_{t=1}^{\infty}$ for an output process Y_1^{∞} based on a filter $\Phi(\theta)$ via the definition

$$\epsilon_t(\theta) = y_t - \hat{y}_t(\theta) \quad (t \in \mathbf{Z}^+),$$

we see that we might as well obtain this process $E_1^{\infty}(\theta)$ via direct linear filtering of Y_1^{∞} with the so-called prediction error filter $P(\theta)$, defined via

$$P(\theta): \quad \begin{cases} \hat{x}_{t+1}(\theta) = [A(\theta) - B(\theta)C(\theta)]\hat{x}_t(\theta) + B(\theta)y_t \\ \epsilon_t(\theta) = -C(\theta)\hat{x}_t(\theta) + y_t \end{cases} \qquad (t \in \mathbf{Z}^+), \quad \hat{x}_1(\theta) = 0,$$

with $\theta \in \Theta$.
Comparing the structure of the equations determining $P(\theta)$ as an i/o–mapping with those of $S(\theta)$, we can represent $P(\theta)$ by the matrix triple $([A(\theta) - B(\theta)C(\theta)], B(\theta), -C(\theta))$. We then have the following Lemmas and Theorems justifying the set–up above.

Lemma A.3 *If the triple $S(\theta) = (A(\theta), B(\theta), C(\theta))$ is minimal, then so is the triple $P(\theta) = ([A(\theta) - B(\theta)C(\theta)], B(\theta), -C(\theta))$ and conversely.*

Proof This is a basic lemma that can be found in many textbooks. We give its proof because of the fact that certain formulas will be used later on as well.
If we denote the controllability matrix of $S(\theta)$ by $\mathcal{C}^{(1)}$ and its observability matrix by $\mathcal{O}^{(1)}$ and likewise denote the controllability matrix of $P(\theta)$ by $\mathcal{C}^{(2)}$ and its observability matrix by $\mathcal{O}^{(2)}$, then these are related by

$$\mathcal{C}^{(2)} = \mathcal{C}^{(1)}T, \quad \mathcal{O}^{(2)} = T^T\mathcal{O}^{(1)},$$

where T denotes the matrix

$$\begin{pmatrix} I & G_1(\theta) & \cdots & G_{n-2}(\theta) & G_{n-1}(\theta) \\ & I & \ddots & & G_{n-2}(\theta) \\ & & \ddots & \ddots & \vdots \\ & O & & I & G_1(\theta) \\ & & & & I \end{pmatrix}$$

in which $G_i(\theta)$ denotes the i–th Markov matrix of $P(\theta)$:

$$G_i(\theta) = \begin{cases} I & \text{for } i = 0 \\ -C(\theta)[A(\theta) - B(\theta)C(\theta)]^{i-1}B(\theta) & \text{for } i > 0. \end{cases}$$

Noticing that T is non–singular for each θ and that minimality follows from rank conditions on the controllability and observability matrices, we obtain the lemma. □

It is worthwhile noticing that $P(\theta)$ can be considered as the inverse system for $S(\theta)$ for each $\theta \in \Theta$: the prediction errors resulting after application of first $S(\theta)$ and then $P(\theta)$ are identical to the white noise input. This a consequence of the fact that the system equations of $S(\theta)$ can be rewitten, by substitution of the expression for $y_t(\theta)$, to yield the prediction error filter equations of $P(\theta)$.
When introducing the Markov parameters $H_i(\theta)$ for $(A(\theta), B(\theta), C(\theta))$ via

$$H_i(\theta) = \begin{cases} I & \text{for } i = 0 \\ -C(\theta)A(\theta)^{i-1}B(\theta) & \text{for } i > 0, \end{cases}$$

we can easily derive, using induction, the following relation to hold:

$$\sum_{k=0}^{s} H_{s-k}(\theta)G_k(\theta) = \begin{cases} I & \text{for } s = 0 \\ 0 & \text{for } s > 0 \end{cases}$$

This means that the inverse of the transfer function of $S(\theta)$ is equal to the transfer function of $P(\theta)$, as the Markov matrices are the coefficients of the (matrix–)Taylor series expansions of the corresponding transfer functions.

Lemma A.4 *$\Phi(\theta)$ is the Kalman filter for $S(\theta)$, for each $\theta \in \Theta$.*

Proof The result is an immediate consequence of the foregoing discussion, as the Kalman filter for a system $S(\theta)$ is characterized by its property of yielding the best linear predictor for y_{t+1} based on all past observations $y_1, \ldots, y_t$. Since the prediction errors resulting from $P(\theta)$ (which is in direct correspondence with $\Phi(\theta)$) are identical to the innovations, we see that $\Phi(\theta)$ is optimal. □

Let us now introduce the following criterion function $V_t(\theta)$ at time $t \in \mathbf{Z}^+$ with respect to the prediction error filters $P(\theta)$, $\theta \in \Theta$, that are applied to a given output process $Y_1^\infty = Y_1^\infty(\theta_*)$ generated by $S(\theta_*)$, where θ_* is unknown:

$$V_t(\theta) = \frac{1}{2}\mathbf{E}_{\theta_*}\{\epsilon_t(\theta)^T\epsilon_t(\theta)\}$$

We have the following result.

Theorem A.5 *At $\theta = \theta_*$ the criterion $V_t(\theta)$ assumes a globally minimal value. For $t > 2n$ this minimum is unique.*

Proof The first part of this theorem is a direct consequence of the fact that the Kalman filter for $S(\theta_*)$ is given by $\Phi(\theta_*)$, leading to optimal prediction errors at each time t identical to the innovations. More formally, we can prove this result via direct calculation proceeding along the following lines.
We can describe the effect of the white noise input v_t on the prediction errors $\epsilon_t(\theta)$ by "connecting" $S(\theta_*)$ and $P(\theta)$, via elimination of the output process Y_1^∞:

$$\begin{cases} \begin{pmatrix} x_{t+1} \\ \hat{x}_{t+1}(\theta) \end{pmatrix} = \begin{pmatrix} A(\theta_*) & O \\ B(\theta)C(\theta_*) & [A(\theta) - B(\theta)C(\theta)] \end{pmatrix} \begin{pmatrix} x_t \\ \hat{x}_t(\theta) \end{pmatrix} + \begin{pmatrix} B(\theta_*) \\ B(\theta) \end{pmatrix} v_t \\ \epsilon_t(\theta) = \begin{pmatrix} C(\theta_*) & -C(\theta) \end{pmatrix} \begin{pmatrix} x_t \\ \hat{x}_t(\theta) \end{pmatrix} + v_t \end{cases}$$

for $t \in \mathbf{Z}^+$ and with $x_1 = 0$, $\hat{x}_1(\theta) = 0$.
We introduce the notation $P_x(t,\theta)$ to describe the covariance of the (extended) state process at time $t \in \mathbf{Z}^+$:

$$P_x(t,\theta) = \mathbf{E}_{\theta_*}\left\{ \begin{pmatrix} x_t \\ \hat{x}_t(\theta) \end{pmatrix} \begin{pmatrix} x_t \\ \hat{x}_t(\theta) \end{pmatrix}^T \right\}$$

As the extended state at time t is fully determined by the stochastic inputs v_s up to time $s = t-1$ and since $\{v_t\}_{t=1}^\infty$ is a white noise process, we have that v_t is independent of the extended state at time t. Thus we obtain the recursion

$$P_x(t+1,\theta) = \tilde{A}(\theta)P_x(t,\theta)\tilde{A}(\theta)^T + \tilde{B}(\theta)\Sigma\tilde{B}(\theta)^T$$

for which we introduce the notation

$$\tilde{A}(\theta) = \begin{pmatrix} A(\theta_*) & O \\ B(\theta)C(\theta_*) & [A(\theta) - B(\theta)C(\theta)] \end{pmatrix}$$

$$\tilde{B}(\theta) = \begin{pmatrix} B(\theta_*) \\ B(\theta) \end{pmatrix}$$

$$\tilde{C}(\theta) = \begin{pmatrix} C(\theta_*) & -C(\theta) \end{pmatrix}$$

This leads to the following general expression for $P_x(t,\theta)$

$$P_x(t,\theta) = \tilde{A}(\theta)^{t-1}P_x(1,\theta)[\tilde{A}(\theta)^T]^{t-1} + \sum_{s=1}^{t-1} \tilde{A}(\theta)^{s-1}\tilde{B}(\theta)\Sigma\tilde{B}(\theta)^T[\tilde{A}(\theta)^T]^{s-1}$$

where a sum over an empty set vanishes.
From the fact that the initial extended state is chosen in a deterministic way as zero, we have that $P_x(1,\theta) = 0$. This leads for $t = 2, 3, \ldots$ to the expression

$$P_x(t,\theta) = \sum_{s=1}^{t-1} \tilde{A}(\theta)^{s-1}\tilde{B}(\theta)\Sigma\tilde{B}(\theta)^T[\tilde{A}(\theta)^T]^{s-1}$$

The criterion value at θ can now be expressed as

$$V_t(\theta) = \frac{1}{2}\mathbf{E}_{\theta_*}\{\epsilon_t(\theta)^T\epsilon_t(\theta)\} = \frac{1}{2}\mathrm{tr}\left\{\mathbf{E}_{\theta_*}\{\epsilon_t(\theta)\epsilon_t(\theta)^T\}\right\} =$$

$$= \frac{1}{2}\mathrm{tr}\left\{\tilde{C}(\theta)P_x(t,\theta)\tilde{C}(\theta)^T\right\} + \frac{1}{2}\mathrm{tr}\{\Sigma\}$$

This shows $\frac{1}{2}\mathrm{tr}\{\Sigma\}$ to be an underbound for the criterion value $V_t(\theta)$.
The formula above can be expressed in terms of $\tilde{A}(\theta)$, $\tilde{B}(\theta)$ and $\tilde{C}(\theta)$ via substitution of the formula for $P_x(t,\theta)$

$$V_t(\theta) = \frac{1}{2}\mathrm{tr}\{\Sigma\} + \frac{1}{2}\mathrm{tr}\left\{\sum_{s=1}^{t-1}\tilde{C}(\theta)\tilde{A}(\theta)^{s-1}\tilde{B}(\theta)\Sigma\tilde{B}(\theta)^T[\tilde{A}(\theta)^T]^{s-1}\tilde{C}(\theta)^T\right\}$$

Using the block partitioning of $\tilde{A}(\theta)$, $\tilde{B}(\theta)$ and $\tilde{C}(\theta)$ we can write

$$\tilde{C}(\theta)\tilde{A}(\theta)^{s-1}\tilde{B}(\theta) = \sum_{k=0}^{s} H_{s-k}(\theta_*)G_k(\theta)$$

with $H_k(\theta)$ and $G_k(\theta)$ as defined before.
Evaluating at $\theta = \theta_*$ we obtain the relation (as above)

$$\sum_{k=0}^{s} H_{s-k}(\theta_*)G_k(\theta_*) = \begin{cases} I & \text{for } s = 0 \\ 0 & \text{for } s > 0 \end{cases}$$

Therefore, since the term for $s = 0$ does not occur in the formula for $V_t(\theta)$, we get

$$V_t(\theta_*) = \frac{1}{2}\mathrm{tr}\{\Sigma\}$$

This proves the minimality of $V_t(\theta)$ at $\theta = \theta_*$.
The second part of the theorem can be shown by establishing its relation to the *partial realization problem*. By reconsidering the general expression in terms of $\tilde{A}(\theta)$, $\tilde{B}(\theta)$ and $\tilde{C}(\theta)$ for $V_t(\theta)$ we get that a global minimum occurs if and only if

$$\tilde{C}(\theta)\tilde{A}(\theta)^{s-1}\tilde{B}(\theta) = 0$$

for $s = 1, 2, \ldots, t-1$. This implies that we must have

$$\sum_{k=0}^{s} H_{s-k}(\theta_*)G_k(\theta) = 0$$

for $s = 1, 2, \ldots, t-1$.
Now remark that by definition $H_0(\theta_*) = I$ and $G_0(\theta) = I$. We therefore can rewrite the set of equations as one (block partitioned) matrix equation:

$$\begin{pmatrix} I & & & & \\ H_1(\theta_*) & I & & O & \\ \vdots & \ddots & \ddots & & \\ H_{t-3}(\theta_*) & & \ddots & I & \\ H_{t-2}(\theta_*) & H_{t-3}(\theta_*) & \cdots & H_1(\theta_*) & I \end{pmatrix} \begin{pmatrix} G_1(\theta) \\ G_2(\theta) \\ \vdots \\ G_{t-2}(\theta) \\ G_{t-1}(\theta) \end{pmatrix} = - \begin{pmatrix} H_1(\theta_*) \\ H_2(\theta_*) \\ \vdots \\ H_{t-2}(\theta_*) \\ H_{t-1}(\theta_*) \end{pmatrix}$$

As a result of the non–singularity of the matrix on the left we see that the matrices $G_1(\theta),\ldots,G_{t-1}(\theta)$ are completely determined by θ_*. Because we have shown that a solution is given by $\theta = \theta_*$ we obtain the necessary and sufficient condition for a global minimum of $V_t(\theta)$

$$G_k(\theta) = G_k(\theta_*) \qquad \text{for } k = 1,\ldots,t-1$$

The problem of realizing matrices $A(\theta)$, $B(\theta)$ and $C(\theta)$ leading to $G_k(\theta)$ with this property is seen to be exactly the partial realization problem with repect to the inverse system $P(\theta_*)$ of $S(\theta_*)$. From standard results from this field, cf. Hazewinkel [24], we then can conclude that if $t > 2n$ the solution is unique, so that by the conditions on the parametrization at hand (in particular the one stating that two i/o–equivalent systems within our class are necessarily specified by the same vector θ) we have that $\theta = \theta_*$ then constitutes a unique global minimum. This completes the proof of the theorem. □

In addition to this proof we remark that *generically* it is sufficient, for uniqueness of the global minimum, to know $G_1(\theta_*),\ldots,G_s(\theta_*)$ with $s = \lceil \frac{2n}{p} \rceil$ denoting the smallest integer greater than or equal to $\frac{2n}{p}$. Therefore, uniqueness is generically obtained for $t > \lceil \frac{2n}{p} \rceil$.
A second remark concerns the necessity of incorporating knowledge of the initial state of the data generating system into the problem structure. The advantage of doing so becomes clear from the expression for $P_x(t,\theta)$ which then no longer depends on the initial state. Letting x_1 be random, e.g. with steady–state covariance, would lead to a situation where θ_* in general does *not* constitute a global minimum for $V_t(\theta)$ as a result of the initial effects. For this, notice that one will always start the prediction error filter with $\hat{x}_1(\theta) = 0$ if no additional information about the initial state of $S(\theta_*)$ is available. Therefore, $P_x(t,\theta)$ will not be stationary. An alternative solution is offered by letting time t run from minus infinity and assuming stationarity. This situation can be obtained in the present set–up by considering $t \to \infty$. Notice that then the initial effects occurring for a random initial state x_1 will vanish. This case is studied in Hanzon [16,17].

References

[1] R.A. Abraham, J.E. Marsden, *Foundations of Mechanics* (2nd ed.). Reading, Mass.: Benjamin & Cummings, 1978.

[2] S. Amari, *Differential-Geometrical Methods in Statistics*, Lecture Notes in Statistics 28. Berlin: Springer–Verlag, 1985.

[3] S. Amari, Differential Geometry of a Parametric Family of Invertible Linear Systems, *Mathematical Systems Theory* 20, 53–82, 1987.

[4] B.D.O. Anderson and J.B. Moore, *Optimal Filtering*. Englewood Cliffs: Prentice–Hall, 1979.

[5] C. Atkinson and A.F.S. Mitchell, Rao's distance measure, *Sankhyā* 43, Series A, 345–365, 1981.

[6] Y. Bard, *Nonlinear Parameter Estimation*. New York: Academic Press, 1974.

[7] W.M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*. New York: Academic Press, 1975.

[8] J.M.C. Clark, The consistent selection of parametrizations in system identification, *Proc. Joint Automatic Control Conference (JACC)*, 576–580. Lafayette, Ind.: Purdue University, 1976.

[9] J.E. Dennis, Jr. and R.B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs: Prentice–Hall, 1983.

[10] R.A. Fisher, *Statistical Methods for Research Workers* (first edition 1925, eleventh edition 1950). Edinburgh: Oliver and Boyd, 1950.

[11] D. Gabay, Minimizing a Differentiable Function over a Differentiable Manifold, *J. of Optimiz. Th. and Appl.* 37, 177–219, 1982.

[12] F.R. Gantmacher, *The Theory of Matrices*, Vol. I and II. New York: Chelsea, 1959.

[13] K.F. Gauss, Theoria Motus Corporum Coelestium, in: *Werke* 7, 240–254, 1809.

[14] K. Glover and J.C. Willems, Parametrizations of Linear Dynamical Systems: Canonical Forms and Identifiability, *IEEE Trans. on Autom. Contr.*, Vol. AC-19, 640–646, 1974.

[15] E.J. Hannan and M. Deistler, *The Statistical Theory of Linear Systems*. New York: John Wiley and Sons, 1988.

[16] B. Hanzon, On a Gauss–Newton identification method that uses overlapping parametrizations, *IFAC Identification and System Parameter Estimation 1985, York, UK*, 1671–1676, 1985.

[17] B. Hanzon, On a coordinate free prediction error algorithm for system identification, in: C.I. Byrnes and A. Lindquist (eds), *Modelling, Identification and Control*. Amsterdam: North–Holland, 595–604, 1986.

[18] B. Hanzon, Riemannian geometry on families of linear systems, the deterministic case. Submitted to: *Math. Contr. Sign. Syst.*

[19] B. Hanzon, *Identifiability, Recursive Identification and Spaces of Linear Dynamical Systems*, CWI Tracts 63, 64. Amsterdam: Centre for Mathematics and Computer Science, 1989.

[20] B. Hanzon, On the differentiable manifold of fixed order stable linear systems, *Systems & Control Letters* 13, 345–352, 1989.

[21] B. Hanzon and R. Ober, Overlapping block–balanced canonical forms and parametrizations: the stable SISO case. Submitted to the 31st CDC, Tucson, Arizona, 1992.

[22] A.C. Harvey, *The Econometric Analysis of Time Series.* Oxford: Philip Allan Publ. Ltd., 1981.

[23] M. Hazewinkel, Moduli and Canonical Forms for Linear Dynamical Systems II: The Topological Case, *Mathematical Systems Theory* 10, 363–385, 1977.

[24] M. Hazewinkel, On the (internal) symmetry groups of linear dynamical systems, in: P. Kramer and M. Dal-Cin (eds), *Groups, Systems and Many Body Physics*, Vieweg Tracts in Pure and Applied Physics 4, 362–404, 1980.

[25] M. Hazewinkel and R.E. Kalman, On invariants, canonical forms and moduli for linear constant finite dimensional dynamical systems, in: G. Marchesini and S.K. Mitter (eds), *Proceedings of the International Symposium on Mathematical System Theory, Udine, Italy*, Lecture Notes in Economics and Mathematical Systems 131, 48–60. Berlin: Springer Verlag, 1976.

[26] R.E. Kalman, On partial realizations, transfer functions and canonical forms, *Acta Polytechnica Scandinavica*, Vol. Ma 31, 9–32, 1979.

[27] P.S. Krishnaprasad, *Geometry of Minimal Systems and the Identification Problem*, Ph.D. Thesis. Cambridge, Mass.: Harvard University, 1977.

[28] S. Kullback, *Information Theory and Statistics.* New York: Wiley, 1959.

[29] S. Kullback and R.A. Leibler, On information and sufficiency, *Ann. Math. Statist.* 22, 79–86, 1951.

[30] H.J. Kushner and D.S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* New York: Springer Verlag, 1978.

[31] C.L. Lawson and R.J. Hanson, *Solving Least Squares Problems.* Englewood Cliffs: Prentice–Hall, 1974.

[32] A. Lichnewsky, Une méthode de gradient conjugué sur des variétés; application à certains problèmes de valeurs propres non linéaires, *Numer. Funct. Anal. and Optimiz.*, Vol. 1, 515–560, 1979.

[33] A. Lichnewsky, *Minimisation des Fonctionnelles Définies sur une Variété par la Méthode du Gradient Conjugué*, Thèse de Doctorat d'Etat. Paris: Université de Paris–Sud, 1979.

[34] L. Ljung, *System Identification: Theory for the User.* Englewood Cliffs: Prentice–Hall, 1987.

[35] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification.* Cambridge, Mass.: MIT Press, 1983.

[36] D.G. Luenberger, The Gradient Projection Method along Geodesics, *Management Science* 18, 620–631, 1972.

[37] A.J.M. van Overbeek and L. Ljung, On–line Structure Selection for Multivariable State Space Models, Report LiTH-ISY-I-0393. Linköping: Linköping University, 1980.

[38] A.J.M. van Overbeek and L. Ljung, On–line Structure Selection for Multi–Variable State–Space Models, *Automatica* 18, 529–543, 1982.

[39] R.L.M. Peeters, Identification on a Manifold of Systems, *Series Research Memoranda* 1992–7. Amsterdam: Free University, FEWEC, 1992.

[40] R.L.M. Peeters, *Ph.D. Thesis.* Forthcoming.

[41] G. Picci, Some Numerical Aspects of Multivariable Systems Identification, *Mathematical Programming Study* 18, 76–101, 1982.

[42] C.R. Rao, Methods of scoring linkage data giving the simultaneous segregation of three factors, *Heredity* 4, 37–59, 1950.

[43] C.R. Rao, *Linear Statistical Inference and Its Applications* (second edition). New York: John Wiley and Sons, 1973.

[44] Yu.A. Rozanov, *Stationary Random Processes.* San Francisco: Holden–Day, 1967.

[45] H.L. Seal, The historical development of the Gauss linear model, *Biometrika* 54, 1–24, 1967.

[46] T. Söderström and P. Stoica, *System Identification.* New York: Prentice–Hall, 1989.

[47] H. Theil, *Principles of Econometrics.* New York: John Wiley and Sons, 1971.

1992-1	R.J. Boucherie N.M. van Dijk	Local Balance in Queueing Networks with Positive and Negative Customers
1992-2	R. van Zijp H. Visser	Mathematical Formalization and the Analysis of Cantillon Effects
1992-3	H.L.M. Kox	Towards International Instruments for Sustainable Development
1992-4	M. Boogaard R.J. Veldwijk	Automatic Relational Database Restructuring
1992-5	J.M. de Graaff R.J. Veldwijk M. Boogaard	Why Views Do Not Provide Logical Data Independence
1992-6	R.J. Veldwijk M. Boogaard E.R.K. Spoor	Assessing the Software Crisis: Why Information Systems are Beyond Control
1992-7	R.L.M. Peeters	Identification on a Manifold of Systems
1992-8	M. Miyazawa H.C. Tijms	Comparison of Two Approximations for the Loss Probability in Finite-Buffer Queues
1992-9	H. Houba	Non-Cooperative Bargaining in Infinitely Repeated Games with Binding Contracts
1992-10	J.C. van Ours G. Ridder	Job Competition by Educational Level
1992-11	L. Broersma P.H. Franses	A model for quarterly unemployment in Canada
1992-12	A.A.M. Boons F.A. Roozen	Symptoms of Dysfunctional Cost Information Systems
1992-13	S.J. Fischer	A Control Perspective on Information Technology
1992-14	J.A. Vijlbrief	Equity and Efficiency in Unemployment Insurance
1992-15	C.P.M. Wilderom J.B. Miner A. Pastor	Organizational Typology: Superficial Foursome of Organization Science?
1992-16	J.C. van Ours G. Ridder	Vacancy Durations: Search or Selection?
1992-17	K. Dzhaparidze P. Spreij	Spectral Characterization of the Optional Quadratic Variation Process
1992-18	J.A. Vijlbrief	Unemployment Insurance in the Netherlands, Sweden, The United Kingdom and Germany
1992-19	J.G.W. Simons	External Benefits of Transport